GeoCLEF 2006: Cross-linguales geographisches Information Retrieval

Kerstin Bischoff, Thomas Mandl, Christa Womser-Hacker
Informationswissenschaft, Universität Hildesheim
Marienburger Platz 22
D-31141 Hildesheim, Deutschland
{mandl, womser}@uni-hildesheim.de

Abstract

Der speziellen Behandlung geographischer Suchanfragen wird im Information Retrieval zunehmend mehr Beachtung geschenkt. So gibt der vorliegende Artikel einen Überblick über aktuelle Forschungsaktivitäten und zentrale Problemstellungen im Bereich des geographischen Information Retrieval, wobei speziell auf das Projekt GeoCLEF im Rahmen der cross-lingualen Evaluierungsinitiative CLEF eingegangen wird. Die Informationswissenschaft der Universität Hildesheim hat in diesem Projekt sowohl organisatorische Aufgaben wahrgenommen als auch eigene Experimente durchgeführt. Dabei wurden die Aspekte der Verknüpfung von Gewichtungsansätzen mit Booleschem Retrieval sowie die Gewichtung von geographischen Eigennamen fokussiert. Anhand erster Interpretationen der Ergebnisse und Erfahrungen werden weiterer Forschungsbedarf und zukünftige, eigene Vorhaben wie die Überprüfung von Heuristiken zur Query-Expansion aufgezeigt.

1 Einleitung

Häufig werden Informationen nicht nur zu einem speziellen Thema, sondern auch mit Bezug auf eine bestimmte geographische Region gesucht. Die (Weiter-) Entwicklung von Geographischen Informationssystemen (GIS), in denen raumbezogene Daten in strukturierter Form gespeichert werden, speziell räumlich abfragt und als Karten visualisiert werden können, weist daher bereits eine lange Tradition auf. Prominente Anwendungen in diesem Bereich sind Fachinformationssysteme bspw. für Umweltdaten, Verkehrs- und Routenplanung oder die digitalen Gelben Seiten. Weniger alt ist die Forschung im Bereich des Geographischen Information Retrieval (GIR), dem Zugänglichmachen und Auffinden von geographisch referenzierten Informationen aus unstrukturierten Daten wie Webdokumenten.

Erste Forschungsarbeiten untersuchen in diesem Kontext die Angemessenheit von erprobten textbasierten IR-Techniken, deren Erweiterungsmöglichkeiten durch externe Wissensressourcen sowie die Potentiale des Einsatzes räumlicher Indexierungs- oder Retrieval-Methoden (z.B. basierend auf Koordinaten zu Längen- und Breitengrad). Problemstellungen sind dabei u.a. die Erkennung und Disambiguierung von geographischen Eigennamen (z.B. *Washington*), die Ähnlichkeitsbestimmung bei vagen Anfragen (z.B. "*in der Nähe von*", *Norddeutschland*) und die Visualisierung der Ergebnisse.

Im Rahmen des Cross Language Evaluation Forum (CLEF) evaluiert das Projekt GeoCLEF GIR-Systeme darüber hinaus auch unter dem Gesichtspunkt der Mehrsprachigkeit, dem Umgang mit geographischen Informationen, die in unstrukturierten Dokumenten verschiedener Sprachen vorliegen. Denn gerade im Hinblick auf die mehrsprachigen Informationen im World Wide Web entsteht hier Mehrwert, wenn monolinguale Anfragen des Benutzers mithilfe maschineller Übersetzungstechniken auch relevante Dokumente anderer Sprachen liefern.

Vieles spricht dabei für die zentrale Rolle von Eigennamen und deren adäquaten Behandlung im Cross-Language Information Retrieval (CLIR). So ist auch im Cross-Language GIR anzunehmen, dass geographische Eigennamen eine hohe Diskriminierungsfähigkeit aufweisen und somit entscheidende Information tragen. Eine korrekte Erkennung dieser Eigennamen ist die Voraussetzung, um geeignete Methoden im Übersetzungsprozess anzuwenden, bspw. in Hinsicht auf die Kompositazerlegung (z.B. *Neuengland* => *New England* vs. *new narrow country*). [Womser-Hacker, 2006]

Nach einem kurzen Überblick über aktuelle Forschungaktivitäten zum GIR wird das Projekt GeoCLEF mit seinen Methoden und Ergebnissen vorgestellt. Es werden eigene Experimente in GeoCLEF beschrieben und potentielle Anschlussarbeiten skizziert.

2 Geographisches Information Retrieval

Um den geographischen Bezug einer Anfrage bzw. eines Dokumentes ermitteln zu können, müssen in einem ersten Schritt geographische Eigennamen korrekt erkannt werden (*Geo-Parsing*). Zur Named Entitiy Recognition (NER) existieren grundsätzlich drei Ansätze: listenbasiert, regelbasiert und mittels maschineller Lernverfahren.

Bei dem Abgleich mit vorgefertigten Listen – im Falle geographischer Eigennamen werden dazu geographische Thesauri, so genannte *Gazetteers*[1], genutzt – dürfte es sich um die zuverlässigste Art handeln, Eigennamen in unstrukturiertem Text zu erkennen. Jedoch sind derartige Ressourcen selten ausreichend umfassend, besonders im Hinblick auf Namensvarianten (*Los Angeles, Stadt der Engel*), und selten frei verfügbar. Regelbasierte Verfahren hingegen arbeiten nach intellektuell erstellten gram-

[1] z.B. Getty Thesaurus of Geographic Names: http://www.getty.edu/research/conducting_research/vocabularies/tgn/; GEOnet World Place Names Server: http://earth-info.nga.mil/gns/html/; World-Gazetteer: http://www.world-gazetteer.com

matikalischen Regeln für das Auftreten von Eigennamen in einer bestimmten Sprache. Dabei scheint die Aufgabe für die deutsche Sprache bspw. wesentlich schwieriger als für die englische, da hier alle Nomen groß geschrieben werden und die Wortstellung mehr Freiheiten zulässt [Womser-Hacker, 2006].

Die Problematik der Sprachabhängigkeit teilen maschinelle Lernverfahren, auch wenn der hohe Aufwand für die Modellierung bzw. Anpassung von Regeln an andere Sprachen durch deren automatische Erstellung anhand annotierter Trainingskorpora entfällt. Wegen der Unzulänglichkeiten der einzelnen Verfahren verwenden viele Systeme mehrere Ansätze sequentiell. In GATE, Teil des SPIRIT-Projektes, [Clough, 2005] folgt bspw. auf einen Thesaurusabgleich die Disambiguierung von Termen mit Treffern in mehreren Listen über den grammatikalischen Kontext. Eine Evaluation freizugänglicher NER-Systeme in Mandl et al. [2005] zeigte recht bescheidene Ergebnisse auf.

Hierarchische Gazetteers können im nächsten Schritt auch genutzt werden, um die Anfrage um alternative Namen und Übersetzungen sowie untergeordnete Länder, Städte, etc. zu expandieren. Speziell für die Problematik fehlender Gazetteer-Einträge für ungenaue Regionen (z.B. *Nordschottland*) stellen Clough et al. [2005a] einen Ansatz vor, durch eine Webanfrage mit Trigger-Phrasen (z.B. „*A is a town in x*“) diese über die Termhäufigkeiten der zurückgelieferten Städtenamen zu modellieren und ggfs. zu expandieren.

Eine wichtige Rolle kommt Gazetteers auch im Prozess der rein geographischen Disambiguierung zu. So kann bei der Mehrdeutigkeitsauflösung von Frankfurt der Kontext nach anderen geographischen Eigennamen durchsucht werden, die sich in ihren Hierarchien überlappen (*World → Europe → Hessen → Frankfurt*). Auf ähnliche Weise nutzen Amitay et al. [2004] einen Gazetteer, um den geographischen Schwerpunkt eines Dokuments durch Zuweisung von *Parent-Regions* zu ermitteln. Sind keine kontextuellen Hinweise vorhanden, kann als Standard eine Entscheidung stets für den Kandidaten mit der kürzesten Hierarchie fallen, da eine größere Bekanntheit angenommen werden kann. [Clough, 2005]

Die Eindeutigkeit der erkannten geographischen Eigennamen ist Grundlage für das *Geo-Coding*, der Zuweisung von Geodaten (Koordinaten) zu den Referenzen. Dieser Verarbeitungsschritt, für den abermals Gazetteers als externe Wissensressourcen nötig sind, ermöglicht räumliche Indexierungs- und Retrievalmethoden. So kann bspw. die räumliche Distanz im Koordinatensystem oder der Grad an Überlappung der Minimum Bounding Rectangles zweier Regionen als Maß für die Relevanz herangezogen werden [Frontiera und Larson, 2004; Chaves et al., 2005].

Ziel von GeoCLEF ist es, Ergebnisse aus den Arbeiten zum Geo-Parsing und Geo-Coding bzw. -Matching unter Beachtung der Mehrsprachigkeit zusammenzuführen.

3 GeoCLEF

2005 fand mit GeoCLEF erstmals ein geographischer Track innerhalb der CLEF-Initiative statt. Die Ergebnisse dieses Pilotprojektes zeigten nicht nur die prinzipielle Durchführbarkeit eines solchen Tracks, sondern auch das große Interesse an dem Themenfeld und vor allem den erheblichen Forschungsbedarf.

3.1 Ziele und Methoden

Entsprechend der Infrastruktur von CLEF werden auch in GeoCLEF Techniken und Systeme anhand einheitlicher Anfragen (*Topics*) gegen ein mehrsprachiges Korpus aus Zeitungsartikeln und Meldungen von Nachrichtenagenturen anhand von anschließenden Relevanzbewertungen miteinander verglichen. Die Topics werden dabei parallel für unterschiedliche Sprachen entwickelt, indem möglichst realistische Benutzeranfragen modelliert, recherchiert und dann von Muttersprachlern in die jeweilige Sprache übersetzt werden.

Im Jahre 2005 konnte in GeoCLEF mit 25 Topics in den Ausgangssprachen Deutsch, Englisch, Spanisch und Portugiesisch – monolingual oder bilingual – in der deutschen oder englischen CLEF-Kollektion experimentiert werden. 2006 waren auch monolinguale Versuche (*Runs*) gegen Kollektionen in Spanisch und Portugiesisch möglich und es kamen Topics in Japanisch hinzu. Die nachfolgende Abbildung zeigt ein Topic aus dem diesjährigen Track:

```
<top>
 <num>GC036</num>
 <DE-title>Automobilindustrie rund um das
  Japanische Meer</DE-title>
 <DE-desc>Küstenstädte am Japanischen Meer mit
  Automobilindustrie oder -werken</DE-desc>
 <DE-narr>Relevante Dokumente berichten von
  Automobilindustrie oder -werken in Städten an der
  Küste des Japanischen Meeres (auch Ostmeer (von
  Korea) genannt), einschließlich wirtschaftlicher oder
  sozialer Ereignisse wie geplante Joint Ventures oder
  Streiks. Neben Japan grenzen auch die Länder
  Nordkorea, Südkorea und Russland an das Japanische
  Meer.</DE-narr>
</top>
```

Abb. 1: Beispiel für ein GeoCLEF-Topic

Während im letzten Jahr die Teilnehmer für ihre Experimente Informationen aus den Tags Title (*title*) und Description (*desc*) sowie zusätzlich aus Concept-Tags, in denen die geographischen Entitäten extrahiert vorlagen, verwendeten, galt es 2006, diese Eigennamen automatisch zu erkennen. Speziell sollte die Nützlichkeit zusätzlicher geographischer Information zur Expansion der Anfrage evaluiert werden. Daher waren sowohl Runs mit den Feldern Title, Description zu absolvieren als auch Runs, die zudem unter den gleichen Parametern den Text des Feldes Narrative (*narr*) – meist beinhaltete dieser die Namen der (Bundes-)Länder einer Region – nutzen. Einige Topics zielten weniger auf eine solche Erweiterung als vielmehr auf geeignete Retrievaltechniken ab *(„Städte im Umkreis von 100 km um Frankfurt“)*.

Aus organisatorischer Sicht erwies es sich dabei als durchaus schwierig, anhand der gegebenen Kollektion Topics zu entwickeln, die einerseits realistische Benutzerbedürfnisse abbilden, zugleich geographisch interessant sind und in allen Sprachkollektionen Treffer vorweisen. So scheinen in diesem Zusammenhang bestimmte Mechanismen der Nachrichtenselektion wie bspw. geographische Nähe, sprachliche oder traditionelle Beziehung und wirtschaftliche Bedeutung zu beeinflussen, ob und

wie oft über ein Ereignis einer bestimmten Region in den Zeitungen einer Kollektion berichtet wird. In allen Kollektionen vertretene geographische Referenzen sind daher zumeist bekannte Regionen bspw. Länder [Clough et al., 2005b]. Trotz Beachtung dieser Problematik bei der Genierung der Topics waren in der Relevanzbewertung 2006 zu einigen Topics in bestimmten Sprachen kaum relevante Dokumente vorhanden.

Da die Ergebnisse aller Teilnehmer für GeoCLEF 2006 aktuell noch nicht vorliegen, sollen kurz die wesentlichen Ergebnisse des Tracks von 2005 genannt werden.

3.2 Ergebnisse und Erfahrungen 2005

Bereits 2005 führten nicht vorhandene relevante Dokumente in der deutschen Kollektion dazu, dass die durchschnittlichen Precison-Werte (MAP)[2] für die mono- und bilingualen Runs ins Deutsche sehr schlecht ausfielen, also die Aufgabe unbeabsichtigt schwieriger war. Die meisten Experimente wurden monolingual Englisch eingereicht und erreichten MAPs von Minimum 0.1464 bis Maximum 0.3936. Hingegen reichten MAPs für monolingual Deutsch lediglich von 0.0535 bis 0.2042. Die insgesamt doch bescheidenen Werte deuten auf die Eigenheiten des GIR und die Notwendigkeit von speziellen, geeigneten Methoden hin. [Clough et al., 2005b]

Dabei nutzten die Teilnehmer verschiedenste Techniken, von ‚einfachen' IR-Techniken ohne jeglichen geographischen Ansatz hin zu Matching mittels Geodaten und Verfahren des Natural Language Processing (NLP), um geographische Hinweise aus Anfrage und Dokument zu extrahieren. Die Erkennung von geographischen Named Entities (NEs) mithilfe verschiedener Techniken wurde jedoch von den meisten Teilnehmern angestrebt. Hauptkritik einiger Teilnehmer war, dass die Topics 2005 üblichen adhoc-Anfragen zu ähnlich waren. Eine reine Keyword-Suche schnitt demnach kaum schlechter ab als elaborierte geographische Methoden. [Clough et al., 2005b]

Die besten Ergebnisse gelangen der Universität Berkeley [Gey und Petras, 2005] für monolingual Englisch durch ein austariertes Blind Relevance Feedback (BRF), auch wenn für einige wenige Topics die Werte durch das Hinzufügen von 30 neuen Termen aus den 5 bestgerankten Dokumenten sanken. Während die Verbesserung durch das BRF im Englischen nur moderat war, zeigte sich im Deutschen eine beachtliche Steigerung der MAP um bis zu 72% des Ausgangswertes.

Manuelle Experimente zur Expansion von geographischen Referenzen (bspw. Europa wurde manuell angereichert um die zugehörigen Ländernamen), brachten überraschenderweise schlechte Ergebnisse für Deutsch und Englisch. Wegen dieser sinkenden Precisionwerte folgern Gey und Petras [2005], dass automatische Expansion mittels geographischer Thesauri nicht sehr viel versprechend ist, es sei denn, es wird ein Boolescher Ansatz – UND-Verknüpfung von inhaltlichem Konzept und geographischer Referenz – verfolgt [auch Larson, 2005]. Diesen Ansatz hat die Universität Hildesheim bei ihrer Teilnahme in GeoCLEF 2006 fokussiert.

[2] Die Mean Average Precison (MAP) gibt die durchschnittliche Precision über festgelegte Recall-Level an. Hier für einen Run gemittelt über alle Topics.

4 Eigene Experimente in GeoCLEF 2006

Nachdem GeoCLEF 2005 im Hinblick auf die Expansion geographischer Eigennamen negative oder zumindest mehrdeutige Ergebnisse gezeigt hatte, sollten in unseren Experimenten die Gesichtspunkte Boolesches Retrieval und Gewichtung bei der Nutzung von zusätzlichen geographischen Informationen untersucht werden. Zur Expansion dienen die Informationen des Narratives der Topics. Die automatische Expansion mittels Gazetteer und Wikipedia für nicht in Gazetteers enthaltene geographische Referenzen wurde noch nicht in die eingereichten Versuche eingebunden, wird aber im Anschluss an die Beschreibung der aktuellen Experimente skizziert.

4.1 Beschreibung der Experimente

Aufbauend auf die Versuche von Mandl et al. [2006] basieren die grundsätzlichen Retrievalfunktionen auf dem Lucene-Paket[3], mit dem Lucene-Stemmer für das Deutsche und Snowball-Stemmer für das Englische. Die Übersetzung der Topics für die bilingualen Versuche beruht auf Babelfish[4], Linguatec[5] und FreeTranslation[6]. Durch die Kombination mehrerer Übersetzer sollen einzelne Fehler abgemildert bzw. zusätzlich Synonyme in der Zielsprache hinzugenommen werden.

Für die Erkennung von Eigennamen wird das maschinelle Lerntool Lingpipe[7] eingesetzt, welches anhand eines trainierten Modells Eigennamen identifiziert und in die Kategorien PERSON, LOCATION, ORGANISATION und MISC klassifiziert. Als Modell diente für das Englische das mitgelieferte, an einem englischen Nachrichten-Korpus trainierte, News-Modell. Da für die deutsche Sprache kein Modell angeboten wird, wurde auf das von Mandl et al. [2006] trainierte zurückgegriffen.

Auch der Index der Kollektion enthält Felder für diese Klassen von Eigennamen, so dass innerhalb des BRF-Prozesses gezielt (geographische) Eigennamen bevorzugt zur Anfragereformulierung herangezogen werden können. Als Idee hinter diesem Schritt steht, dass somit aus den topgerankten Dokumenten zu einer geographischen Suche weitere Eigennamen von zugehörigen Städten, Regionen oder Ländern gefunden werden und in die Anfrage eingehen können. Derart ermittelte Geo-Entitäten werden der Anfrage dabei über ein Boolesches UND zugefügt, um auch das inhaltliche Kriterium zu erfüllen. Auch in Versuchen ohne BRF wurden daher erkannte Geo-NEs aus den Feldern Title, Description und Narrative über UND mit dem Inhalt verknüpft. Die generellen Verarbeitungsschritte können daher wie folgt schematisch dargestellt werden, wobei je nach Run nur bestimmte Schritte ausgeführt bzw. Parameter geändert wurden:

Topic → (Übersetzung) → (NER und Gewichtung) → Stoppworttilgung → Stemming → Anfrage Boolesches UND vs. ODER → (BRF ggfs. mit Gewichtung von geographischen Eigennamen)

Abb. 2: Verarbeitungsschritte des Systems

[3] http://lucene.apache.org/java/docs/
[4] http://babelfish.altavista.com/
[5] http://www.linguatec.de/onlineservices/pt
[6] http://www.freetranslation.com/
[7] http://www.alias-i.com/lingpipe/

Sprache	Feld	NEs	BRF	Query	MAP
En	title, desc	-	-	ODER	0,1676
En	title, desc, narr	-	-	ODER	0,1747
En	title, desc	gewichtet	5 docs, 25 terms, NEs + GeoNEs gewichtet	UND	0,1166
En	title, desc, narr	gewichtet	5 docs, 25 terms, NEs + GeoNEs gewichtet	UND	0,1213
En	title, desc	-	5 docs, 20 terms, GeoNEs gewichtet	ODER	0,1875
De	title, desc	-	5 docs, 25 terms	ODER	0,1558
De	title, desc, narr	-	5 docs, 25 terms	ODER	0,1601
De	title, desc	gewichtet	5 docs, 25 terms, NEs + GeoNEs gewichtet	UND	0,1214
De	title, desc, narr	gewichtet	5 docs, 25 terms, NEs + GeoNEs gewichtet	UND	0,1134
De → En	title, desc	-	-	ODER	0,1504
De → En	title, desc, narr	-	-	ODER	0,1903
De → En	title, desc	gewichtet	5 docs, 25 terms, NEs + GeoNEs gewichtet	UND	0,1456
De → En	title, desc, narr	gewichtet	5 docs, 25 terms, NEs + GeoNEs gewichtet	UND	0,1565
De → En	title, desc	-	5 docs, 20 terms, GeoNEs gewichtet	ODER	0,1603
En → De	title, desc	-	5 docs, 25 terms	ODER	0,1186
En → De	title, desc, narr	-	5 docs, 25 terms	ODER	0,1315
En → De	title, desc	gewichtet	5 docs, 25 terms, NEs + GeoNEs gewichtet	UND	0,0969
En → De	title, desc, narr	gewichtet	5 docs, 25 terms, NEs + GeoNEs gewichtet	UND	0,1046

Abb. 3: MAPs der einzelnen Hildesheimer Runs

So wurden für die Aufgaben monolingual Englisch und bilingual Deutsch → Englisch Versuche eingereicht, in denen als Base Run weder NEs erkannt noch BRF durchgeführt wurde, sondern nur die genutzten Felder variiert wurden. In zwei weiteren Runs wurden darüber hinaus NEs erkannt und stärker gewichtet, sowie im BRF das Hinzufügen von geographischen Eigennamen forciert. In einem fünften Versuch – ohne die Informationen des Feldes Narrative – wurde nur auf Geo-NEs im BRF getestet.

Die monolingualen Versuche Deutsch und die bilingualen Versuche Englisch → Deutsch waren analog konzipiert, jedoch entfielen die Runs ohne BRF, da diese im Training an den GeoCLEF-Daten von 2005 die schlechtesten Ergebnisse lieferten.

An dieser Stelle können nur erste Folgerungen aus den Ergebnissen der Teilnahme dargestellt werden, da bislang die eingesetzten Techniken der anderen Teilnehmer und deren Abschneiden[8] unklar sind sowie eine genaue Analyse der Ergebnisse für die einzelnen Topics noch aussteht.

4.2 Erste Ergebnisse

Die Ergebnisse sind über alle Versuche hinweg keineswegs zufrieden stellend, liegen jedoch im Durchschnitt der Experimente aller Teilnehmer. So zeigt sich auch in GeoCLEF 2006, dass die Aufgabe des GIR nicht trivial ist, denn auch die besten Teilnehmer konnten nur eine MAP von 0,3223 für monolingual Englisch, 0,2229 für monolingual Deutsch und 0,1682 für bilingual Englisch → Deutsch verzeichnen.

Trotz der geringen Veränderungen der MAP scheint in unseren Versuchen kein negativer Effekt einer Expansion um untergeordnete geographische Entitäten spürbar – sowohl bei Boolescher Verknüpfung als auch bei einfacher Gewichtung. Unter gleichen Parametern verbesserte sich stets die Precision unter Hinzunahme des Feldes Narrative. Im Falle des besten Laufes für bilingual Deutsch → Englisch dürfte der Anstieg der MAP auf 0,1903 jedoch daran liegen, dass die Erweiterung um den Begriff *Indonesien* das einzige relevante Dokument zu diesem Topic zurückliefern konnte.

Die im internen Vergleich recht guten Ergebnisse der Höhergewichtung von Geo-Entitäten innerhalb des BRF, könnten darauf hindeuten, dass die Art der zusätzlichen geographischen Information in den Narratives nicht unbedingt die für die Expansion geeigneteste sein könnte.

Die Verbesserung der Werte für die Runs mit Boolescher UND-Verknüpfung in der bilingualen Bedingung Deutsch → Englisch ggü. monolingual Englisch durch den kombinierten Einsatz mehrerer Übersetzer ist interessant. Zu prüfen ist, ob dies auf dadurch gewonnene Synonyme oder eine wegen Varianten in der Wortstellung erleichterte NER zurückzuführen ist. Für die Versuche mit Gewichtung und boolescher Verknüpfung von (Geo-)NEs ist eine weitere Analyse der Performanz der NER nötig, bevor Schlussfolgerungen gezogen werden können. Eine erste Durchsicht zeigt, dass einige NEs gerade im Deutschen nicht korrekt erkannt wurden.

Eine genaue Analyse sowohl der Performanz über die unterschiedlichen Topics mit ihren jeweils eigenen Anforderungen (die MAP-Werte für die einzelnen Topics variieren stark) und die Isolierung der Einzeleffekte der eingesetzten Verfahren ist der nächste Schritt, um erfolgreiche(re) Techniken zu ermitteln.

Ausblick

Die niedrigen MAP-Werte zeigen, dass im Anschluss an die GeoCLEF-Teilnahme 2006 besonders im Hinblick auf die Rolle von (geographischen) NEs eine genaue Analyse durchgeführt werden muss. Für weitere Versuche ist daher als Grundlage die Verbesserung der NER durch Fusion verschiedener Ansätze geplant.

Die automatische Expansion mithilfe eines Gazetteers wird umgesetzt, wobei den Heuristiken für eine erfolgreiche Anreicherung spezielle Beachtung geschenkt werden soll. So scheinen Bekanntheit, (wirtschaftliche) Relevanz, geographische Nähe einer Region zu beein-

[8] Es liegen für jede Aufgabe der Mittelwert aller eingereichten Experimente, der beste und schlechteste Wert vor. Diese sind der Email-Kommunikation mit Ray Larson (an die GeoCLEF-Organisatoren) vom 27.07.2006 entnommen.

flussen, ob und wie eine Expansion nötig oder sogar sinnvoll ist. Nicht nur zur Weiterverfolgung der Idee, im BRF für die Expansion geeignete andere geographische Eigennamen zu finden, sind zudem Techniken zur Disambiguierung und zur Ermittlung des geographischen Schwerpunktes eines Dokumentes zu integrieren.

Eine angedachte Anbindung von Wikipedia als externer Wissensressource, durch welche Referenzen ohne Gazetteer-Eintrag (z.B. *Norddeutschland, Warschauer Pakt*) geographisch expandierbar werden sollen, basiert ganz wesentlich darauf. Darüber hinaus ist im Hinblick auf Synonyme eine Erweiterung des Systems z.B. um den Wortschatz Leipzig geplant.

Literatur

[Amitay et al., 2004] Einat Amitay, Nadav Har'El, Ron Sivan und Aya Soffer. Web-a-Where: Geotagging Web content. In *SIGIR 2004: Proceedings of the 27th Annual International ACM SIGIR Conference on Research and Development in Information Retrieval*, Seiten 273-280, Sheffield, Großbritannien, Juli 2004, ACM.

[Chaves et al., 2005] Marcirio Silveira Chaves, Bruno Martins und Mário J. Silva. Challenges and Resources for Evaluating Geographical IR. In *Proceedings of the 2nd International Workshop on Geographic Information Retrieval*, CKIM 2005, Seiten 65-69, Bremen, Deutschland, November 2005.

[Clough, 2005] Paul Clough. Extracting Metadata for Spatially-Aware Information Retrieval on the Internet. In *Proceedings of the 2nd International Workshop on Geographic Information Retrieval*, @CKIM 2005, Seiten 25-30, Bremen, Deutschland, November 2005.

[Clough et al. 2005a] Paul Clough, Hideo Joho und Ross Purves. Identifying imprecise regions for geographic information retrieval using the web. In *Proceedings of the GIS RESEARCH UK 13th Annual Conference*, Seiten 313-318, Glasgow, Großbritannien, 2005.

[Clough et al., 2005b] Paul Clough, Frederic Gey, Hideo Joho, Ray Larson, Vivien Petras und Mark Sanderson. GeoCLEF: the CLEF 2005 Cross-Language Geographic Information Retrieval Track. In *Working Notes for the CLEF 2005 Workshop*, Wien, Österreich, September 2005, http://www.clef-campaign.org/2005/ working_notes/.

[Frontiera und Larson, 2004] Patricia Frontiera und Ray R. Larson. Evaluation and Usability – Spatial Ranking Methods for Geographic Information Retrieval (GIR) in Digital Libraries. In *Proceedings of the 8th European Conference on Research and Advanced Technology for Digital Libraries*, ECDL 2004, Seiten 45-56, Bath, Großbritannien, September 2004, Lecture Notes in Computer Science 3232, Springer.

[Gey und Petras, 2005] Frederic Gey und Vivien Petras. Berkeley2 at GeoCLEF: Cross-Language Geographic Information Retrieval of German and English Documents. In *Working Notes for the CLEF 2005 Workshop*, Wien, Österreich, September 2005, http://www.clef-campaign.org/2005/ working_notes/.

[Goodchild et al., 2005] Michael Goodchild, Paul A. Longley, David J. Maguire und David W. Rhind. *Geographic Information Systems and Science*. John Wiley and Sons, Chichester, 2. aktualisierte Auflage 2005.

[Jones und Purves, 2004] Chris Jones und Ross Purves. *Workshop on Geographic Information Retrieval, SIGIR 2004*. SIGIR Forum, 38(2): 53-56, Dezember 2004.

[Jones und Purves, 2005] Chris Jones und Ross Purves, Herausgeber. *Proceedings of the 2005 Workshop on Geographic Information Retrieval*, GIR 2005, Bremen, Deutschland, November 2005, ACM.

[Larson, 2005] Ray R. Larson. Chesire II at GeoCLEF: Fusion and Query Expansion for GIR. In *Working Notes for the CLEF 2005 Workshop*, Wien, Österreich, September 2005, http://www.clef-campaign.org/2005/ working_notes/.

[Mandl et al., 2005] Thomas Mandl, René Schneider, Pia Schnetzler und Christa Womser-Hacker. Evaluierung von Systemen für die Eigennamenerkennung im cross-lingualen Information Retrieval. In B. Fisseni, Hans-Christian Schmitz, Bernhard Schröder und Petra Wagner, Herausgeber. *Sprachtechnologie, mobile Kommunikation und linguistische Ressourcen: Beiträge zur GLDV Tagung 2005 in Bonn*, [Sprache, Sprechen und Computer/ Computer Studies in Language and Speech 8], Seiten 145-157, Peter-Lang, Frankfurt/Main et al., 2005.

[Mandl et al., 2006] Thomas Mandl, René Schneider und Robert Strötgen. A Fast Forward Approach to Cross-lingual Question Answering for English and German. In Fredric C. Gey, Julio Gonzalo, Gareth J.F Jones, Michael Kluck, Bernardo Magnini, Henning Müller, Carol Peters and Maarten de Rijke, Herausgeber. *Accessing Multilingual Information Repositories: 6th Workshop of the Cross-Language Evaluation Forum*, CLEF 2005, Wien, Österreich, Revised Selected Papers. Lecture Notes in Computer Science 4022, Springer, 2006.

[Womser-Hacker 2006] Christa Womser-Hacker. Zur Rolle von Eigennamen im Cross-Language Information Retrieval. In Ilse Harms, Heinz-Dirk Luckhardt und Hans W. Giessen, Herausgeber. *Information und Sprache. Beiträge zu Informationswissenschaft, Computerlinguistik, Bibliothekswesen und verwandten Fächern*. Festschrift für Harald H. Zimmermann zum 65. Geburtstag, K.G. Saur, München, 2006.